Absoluut en relatief meten, 5 (2) pp. 30-54.

Terwel, J.

published in
Info, Informatiebladen van het Instituut voor Onderwijskunde der RU/Groningen
1973

document version
Publisher's PDF, also known as Version of record

Info

Informatiebladen van het Instituut voor onderwijskunde der R.U. Groningen

Abrobuut en relatief werken

5e jaargang nr.2
november 1973

INFO

Informatiebladen van het Instituut voor Onderwijskunde
der R.U. te Groningen.

Brieven en stukken voor de redaktie dient men te zenden aan:
Drs. E. J. Boerma, Instituut voor Onderwijskunde der R.U.,
Westerhaven 16, Groningen. Telefoon (050) - 115258
of (050) - 114379 (sekretaris).

Abonnement f 16,00 per jaar.
Info verschijnt 6 x per jaar.
Jaargangen lopen van september tot september.

Alle bestellingen en afleveringen geschieden via het Instituut voor
Onderwijskunde.
Betaling per bank c.q. postgiro: t.n.v. Info
A.B.N. Groningen rek.nr. 570047455.
Postgiro A.B.N. 802394.

Inhoud

Absoluut en relatief meten

I N H O U D:

1. INLEIDING

In het kader van een werkkollege didaxologie (DII 72/73) over het onderwerp interne schoolorganisatie, is door één van de werkgroepen een literatuurstudie gemaakt over "Mastery Learning en Aptitude Treatment Interaction".
Het resultaat van genoemde studie vormt het vertrekpunt voor dit literatuur onderzoek. In het bijzonder zal worden verder gewerkt op de onderscheiding absoluut en relatief meten (Lagerweij e.a., 1973, 243)
Het globale verschil tussen beide procedures kan voorlopig als volgt worden omschreven:
Absolute meetprocedures verschaffen informatie over wat een individu kan en niet kan, onafhankelijk van de verhouding van zijn prestatie tot de prestaties van andere individuen.
Relatieve meetprocedures verschaffen informatie over de rangorde van de prestaties van individuen. M.a.w. relatieve metingen geven aan hoeveel leerlingen een bepaalde leerling met zijn prestatie overtreft.
Deze onderscheiding blijkt al vele tientallen jaren onderwerp van discussie te zijn in de Amerikaanse psychometrische literatuur.
De laatste jaren is de belangstelling voor de problematiek van absoluut versus relatief meten, vooral onder invloed van de diskussie over mastery learning, enorm toegenomen.
Tegelijkertijd is hiermee het probleem in een veel bredere kontekst geplaatst. Naast de statistische en psychometrische aspekten worden problemen van onderwijsfilosofische aard expliciet in de diskussie betrokken.
In het volgende worden enkele belangrijke aspekten van deze problematiek benaderd vanuit de recente internationale literatuur.
Hierbij zijn de volgende vragen richtinggevend geweest:

1. Wat is het onderscheid tussen absoluut en relatief meten?
2. Is deze onderscheiding in skoringsprocedures gerelateerd aan een onderscheiding in:
   - toetsdoelen
   - onderwijsleersituaties
   - onderwijsfilosofieën
   - meetinstrumenten
   - meettheorieën
3. Moet er een keuze worden gemaakt tussen absoluut en relatief meten?

2. ABSOLUUT VERSUS RELATIEF METEN: HERKOMST EN DEFINITIE

2.1. Geschiedenis

Bij het toetsen van studieprestaties wordt in Amerika al vanaf 1900 onderscheid gemaakt tussen relatief en absoluut meten (Ebel, 1965, 406). Tot ± 1920 was het gebruikelijk om de prestaties in procenten uit te drukken.
Een student die alle doelstellingen voor een bepaalde onderwijsperiode had bereikt kreeg het cijfer 100 (perfect mastery).
Voor het (theoretische) geval dat een leerling helemaal niets had geleerd werd een nul gegeven.
De cesuur werd gewoonlijk gelegd bij een percentage goede antwoorden tussen 60 en 75%. Deze wijze van prestatiemeting wordt soms gekarakteriseerd als een vorm van absoluut meten.
Deze methode werd na 1920 vervangen door een vorm van relatief meten: het vijf-letter A-B-C-D-F- systeem. De reden voor deze vervanging was o.a. dat het onmogelijk bleek om precies te specificeren wat nu eigenlijk onder perfekte beheersing moest worden verstaan. (Ebel, 1970, 2).

Interessant is verder dat reeds veertig jaar geleden, door prof. H.C.Morrison van de universiteit van Chicago, een strategie voor mastery learning werd ontwikkeld.
Hierbij speelden absolute metingen een belangrijke rol (Ebel, 1970, 7).
De ideeën van Morrison waren aanvankelijk zeer populair maar langzamerhand verloren ze steeds meer terrein.
In de zestiger jaren werd de problematiek opnieuw actueel vooral door een artikel van R.Glaser in de "American Psychologist" (Glaser, 1963). In dit artikel maakt Glaser een onderscheid tussen "criterion-referenced measures" en "norm-referenced measures": "The principal difference between these two kinds of information lies in the standard used as a reference.
What I shall call criterion-referenced measures depend upon an absolute standard of quality, while what I term norm-referenced measures depend upon a relative standard".
Tegelijkertijd stelt Glaser de vraag of de klassieke meettherorie (aptitude measurement) wel bruikbaar is op het gebied van "achievement measurement".

2.2. Algemene definitie: Warries.

In Nederland is de terminologie absoluut versus relatief meten het eerst gebruikt door Warries. Hij geeft de volgende omschrijving:
"Absoluut meten van studieprestaties is die vorm van meten waarbij het cijfer, de score of andere beoordeling van de leerling aangeeft hoever hij gevorderd is op de (wel omschreven) weg van kennisverwerving. Een dergelijk oordeel kan b.v. verwijzen naar de proportie gekende Franse woorden uit een lijst of naar het aantal leerstofeenheden over contemporaine geschiedenis voor de M.A.V.O, dat een leerling met goed resultaat heeft doorgewerkt. "Goed resultaat" kan nader worden gedefinieerd, b.v. door te zeggen dat 90% van de bijbehorende vragen goed beantwoord is.
Bij absoluut meten behoeft de individuele uitkomst niet bepaald te worden door vergelijking met andere leerlingen al blijft een vergelijking natuurlijk wel mogelijk door te kijken hoever de leerling is gekomen in verhouding tot de anderen.
De maatstaf echter voor de bepaling van de meet-uitkomst heeft niet te maken met de prestaties van andere leerlingen op dat moment. De meetschaal voor absoluut meten wordt gevormd door de weg naar kennisverwerving in het betreffende studie gebied of schoolvak zoals beschreven door de deskundigen. De maatstreepjes op dit continuum worden niet zoals bij relatief meten gevormd door percentages medeleerlingen maar door kritische punten gedefinieerd in termen van tussentijdse toetsen die je (bijna) helemaal goed moet doen.
Relatief meten van studieprestaties is die vorm van meten waarbij de leerling pas zijn score cijfer of beoordeling toebedeeld krijgt nadat zijn antwoorden zijn vergeleken met die van de andere leerlingen die hetzelfde werk hebben gemaakt.
Het vergelijkende aspekt van deze beoordelingswijze is bijzonder duidelijk bij klassificerende uitspraken als: "de slechtste van zijn klas" of "behoort tot de beste 10% van zijn jaargroep".
Relatief meten verschaft geen inzicht in de vorderingen van de leerling in termen van de leerstof die hij beheerst maar geeft een vergelijkende uitspraak of een cijfer dat zijn positie in de groep aanduidt. Uit de prestaties van de groepsleden- en dit is

de essentie van relatief meten -wordt de schaal gekonstrueerd de maatstaf waarmee de individuele leerling wordt beoordeeld. Met een zekere prestatie kan een leerling in een slecht presterende groep tot de besten en in een zeer goed presterende groep tot de slechtsten behoren. Dikwijls wordt de relatieve positie van de leerling aangegeven door een percentielskore of met een standaardskore, beide vertaalbaar in het percentage medeleerlingen dat de leerling achter zich heeft gelaten" (Warries 1970, 431).
Warries geeft de voorkeur aan de toepassing van een absolute meetmethode. Hij meent dat absoluut meten vooral geëigend is bij teacher-made tests die door kleine groepen gemaakt worden in het voortgezet onderwijs.
Niet alleen dat dan uit meettechnisch oogpunt de relatieve meting minder geschikt is maar er zijn ook andere bezwaren: neiging tot geforceerde differentiatie en geringe kommunikatie waarde.
Het meest fundamentele bezwaar dat Warries aanvoert tegen relatief meten betreft de filosofie achter deze skoringsprocedure.
Deze onderwijsfilosofie is een soort "ekonomische schaarste filosofie" die als volgt kan worden samengevat:

a. Velen zijn geroepen doch weinigen uitverkoren.
b. Onderlinge wedijver bevordert de prestatiedrang.
c. Een aantal leerlingen zal het nooit leren.
d. Toetsen moeten differentiëren tussen knappen en dommen.
e. Toetsen moeten moeilijk zijn.
f. Op een toets moet de leerling zich niet kunnen voorbereiden (Warries, 1970, 434).

2.3. Operationele definitie: Wijnen

Het onderscheid tussen absoluut en relatief meten wordt door Wijnen zeer verhelderend beschreven in termen van operaties die nodig zijn om te komen tot het bepalen van de cesuur voldoende/onvoldoende (Wijnen, 1971, 20).
Wijnen noemt de volgende operaties:

1. Het kiezen van het referentie punt
2. Het kiezen van de nog juist toelaatbare afstand
   hierbinnen moeten 2 beslissingen worden genomen:

   a. het kiezen van de meeteenheid
   b. het kiezen van de kwantificering van deze meeteenheid.

Bij absoluut meten wordt een absoluut referentiepunt gekozen, vóórdat de studietoets wordt afgenomen.
Als voorbeelden van absolute referentiepunten kunnen worden genoemd: de maximale skore, de toevalsskore, het nulpunt.
Vervolgens kiest men de meeteenheid b.v. procenten of vragen.
Daarna bepaalt men de grensskore in termen van het aantal meeteenheden vanaf het gekozen referentiepunt b.v. 65% vanaf het nulpunt of 10 vragen vanaf de maximum skore.
De redenering achter deze procedure kan als volgt worden omschreven: het proces van kennisverwerving kan theoretisch worden gezien als een proces dat zich uitstrekt op een kontinuum van geen beheersing (no proficiency at all) tot volledige beheersing (perfect performance) (Glaser, 1963, 519).
De individuele prestatie, zoals blijkt uit de toetsverrichting, ligt ergens op een punt op dit kontinuum.
Men kiest nu een bepaald punt (het kriterium) op dit kontinuum

waaraan de individuele prestatie van een leerling wordt afgemeten. Het kriterium is dus een operationalisatie van de doelstellingen die bereikt moeten worden ongeacht de feitelijk bereikte resultaten van de groep.
De methode van Nedelsky voor het bepalen van de cesuur kan in bepaald opzicht worden beschouwd als een vorm van absoluut meten, omdat het kriterium onafhankelijk is van de gegevens die pas na de invulling van de test kunnen worden verkregen.
Dit maakt het mogelijk, de aftestgrans vooraf vast te stellen en bekend te maken.
Bij relatief meten wordt een relatief referentiepunt gekozen nadat de studietoets is afgenomen. Voorbeelden van relatieve referentiepunten zijn: het gemiddelde en de mediaan. Een veel gebruikte meeteenheid bij relatieve systemen is de standaarddeviatie. De grensskore wordt vervolgens bepaald in termen van een x-aantal standaarddeviaties vanaf het gemiddelde of vanaf de mediaan.
"Grading on the curve kan worden getypeerd als een methode met een relatief referentiepunt, met de standaarddeviatie van het gemiddelde als meeteenheid, waarbij de kwantificering van de nog juist toelaatbare afstand gelijk is aan 1,5 standaarddeviatie". (Wijnen, 1971, 37).
Wijnen heeft een zekere voorkeur voor een relatieve meetmethode zowel voor selektiedoeleinden als voor formatieve evaluatie bij mastery testing (Wijnen 1973).
Op de argumentatie kan echter pas worden ingegaan wanneer de psychometrische aspekten aan de orde zijn geweest.
De beslissingsregel die Wijnen voorstelt voor selektiedoeleinden adviseert die kandidaten af te wijzen, van wie de prestatie lager is dan het gemiddelde van de groep verminderd met tweemaal de standaardmeetfout.
Een relatief referentiepunt, het rekening houden met de kwaliteit van de meting en een gebruikelijke kwantificering van de nog juist toelaatbare afstand zijn daardoor in één methode opgenomen (Wijnen, 1971, 58)

2.4. De illusie van het absolute oordeel.

Bij de definiëring is min of meer als vanzelfsprekend aangenomen dat er een duidelijk onderscheid bestaat tussen absoluut en relatief meten.
Bij nader inzien moet de genoemde onderscheiding eerder worden beschouwd als een ideaal-typische konstruktie dan als een praktische realiteit.
De strikt absolute beoordeling blijkt in feite een illusie te zijn. Beoordelingsprocedures waarbij de norm vooraf wordt bepaald (dus zonder dat de aktuele prestatie van de te beoordelen groep een rol speelt) zijn in werkelijkheid niet zo absoluut als veelal wordt aangenomen.
In de praktijk blijkt dat zogenaamde "absolute" normen worden bepaald in relatie tot de gemiddelde prestatie van een abstrakte of een bestaande referentiegroep.
De twijfel aan de mogelijkheid van het absolute oordeel is bij enkele auteurs "tussen de regels door" te lezen: bij Wijnen's bespreking van de "absolute" methode van Nedelsky (Wijnen, 1971, 32) en bij Ebel's uiteenzetting over het Amerikaanse beoordelingssysteem zoals dat werd gehanteerd vóór 1920. (Ebel, 1965, 406).

In een recent artikel van Hofstee over norm-handhaving bij toetsen komt dit probleem meer expliciet aan de orde.
Hofstee stelt: "dat alle normen uiteindelijk (of liever: in eerste instantie) niet anders dan relatief kunnen zijn. Ze moeten ooit zijn bepaald aan de hand van de prestaties van een of andere "eerste groep".
Wanneer géén geoefend typiste meer dan vijf woorden per minuut zou typen, was er geen denken aan dat de norm zou liggen waar hij nu ligt. De tegenwerping dat docenten in nieuwe situaties wel degelijk bij voorbaat normen kunnen stellen en dat ook doen, kan gemakkelijk worden ontkracht.
Wat in zo'n geval plaatsvindt is een generalisatie vanuit reeds bekende situaties (op grond van voorkennis over de disposities van de leerlingen en de aard van de stof).
Als die generalisatie een akseptabel resultaat oplevert - en de tolerantiegrenzen daarvoor plegen zeer ruim te zijn - ontstaat al snel de illusie dat van een absoluut oordeel kan worden gesproken. In feite echter is de mens de maat van alle dingen.
De "**eerste** normbepaling" is dus altijd gebaseerd op de frekwentieverdeling van de prestaties van een bepaalde groep, bestaand of bedacht". (Hofstee, 1973, 216).
Het is bijzonder interessant dat op het gebied van de groepsdynamika deze problematiek in meer algemene termen wordt beschreven.
In zijn bespreking van de verschillende soorten referentiegroepen stelt Thelen m.b.t. de abstracted group:

"The most usual relic of the abstracted group is the value system, remembered long after times, places, circumstances, and individuals are forgotten. In our quest for authority and certainty we often carry over to the actual group the values and attitudes that made sense for some previous group; we hope that they will guide us out of the perplexities of the present group.
When values abstracted from these forgotten groups or cultures come into conflict with the evaluative criteria of a current actual group, the former had better be examined carefully to see if there is enough similarity of circumstance that they can be accepted as the "absolutes" one wishes they were".(Thelen, 1970, 232)

Bovenstaande groepsdynamische uiteenzetting van Thelen loopt parallel aan de statistische argumentatie van Hofstee.
Dit wordt vooral duidelijk als Hofstee het heeft over " ... de pretenties van beoordelaars die denken absoluut te kunnen oordelen: zij doen dat vanuit hun Oneindige Ervaring" (Hofstee, 1973, 217).
Zowel Thelen als Hofstee zijn van mening, dat het onjuist is om aan de normen (resp. prestaties) van de "abstracted group" (resp. "Eerste Groep") een overheersende waarde toe te kennen en de normen (resp. prestaties) van de aktuele groep buiten beschouwing te laten.
Op de ethische implikaties van het bovenstaande kan in het kader van dit referaat niet worden ingegaan.
Wel kan worden opgemerkt dat Hofstee wellicht ten onrechte Protagoras citeert.
Het is namelijk niet waarschijnlijk dat Protagoras een soort (statistisch) gemiddelde heeft bedoeld met zijn uitspraak:
De mens is de maat van alle dingen, van het zijnde voor het zijn, van het niet-zijnde voor het niet zijn. (Störig, 1966, 139) (Bakker, 1969, 26).

3. ABSOLUUT VERSUS RELATIEF METEN: PSYCHOMETRISCHE ASPEKTEN

3.1. Verkenning

Het verschil tussen absoluut en relatief meten is in het voorgaande voornamelijk gedefinieerd in termen van scoringsprocedures.
Diverse auteurs wijzen echter tevens op verschillen in toetskonstruktie en meettheorieën.
Op dit punt blijkt de problematiek erg gekompliceerd te zijn.
Dit wordt nog versterkt door verschillende meningen vooral t.a.v. een al of niet noodzakelijk geacht verschil in meettheorie.
Deze onzekerheid vormt een nieuw element in de diskussie rondom absoluut en relatief meten.
Tot voor kort werd de toets als een gegeven beschouwd. De vraag was: hoe bepalen we een zinvolle score. M.a.w. er bestond een grote mate van overeenstemming inzake het konstrueren en analyseren van studietoetsen (Popham en Husek, 1969, 1).
Momenteel wordt de bruikbaarheid van de klassieke testtheorie voor bepaalde vormen van prestatiemeting ter diskussie gesteld.
Warries wijst op de noodzaak van een bewustwordingsproces bij psychologen, dat zou kunnen leiden tot het inzicht dat de klassieke testtheorie wel eens minder bruikbaar zou kunnen zijn voor het toetsen van prestaties in het onderwijs (Warries, 1970).
Elders pleit Warries voor verschillende benaderingen inzake konstruktie, afname, scoring en analyse van studietoetsen afhankelijk van het gestelde doel. (Warries, 1971).
Ook Nuy is van mening dat, afhankelijk van het gebruiksdoel (selektie of diagnose) niet alleen de scoringsprocedure maar ook het meetinstrument en de psychometrische analyse-technieken verschillend moeten zijn.
Ten behoeve van selektie zal men studieprestaties relatief meten met norm-referenced toetsen.
Ten behoeve van diagnose zal men studieprestaties absoluut meten met criterion-referenced toetsen (Nuy, 1973, 651).
Het verschil tussen een criterion-referenced toets en een norm-referenced toets is, volgens Popham en Husek, op het eerste gezicht niet aan te geven. Een criterion-referenced test zou ook gebruikt kunnen worden als norm-referenced test.
Hoewel men zich het omgekeerde moeilijker kan voorstellen (Popham en Husek, 1969, 2).
Popham en Husek wijzen op de noodzaak van het ontwikkelen van een geheel nieuw theoretisch kader: een criterion-referenced approach. Een nieuwe benadering inzake toetskonstruktie, validering en betrouwbaarheidsbepaling.
Daarentegen stelt S.A. Livingston dat de principes van de klassieke testtheorie bruikbaar zijn bij de betrouwbaarheidsbepaling van criterion-referenced test.
Livingston ziet norm-referenced measurement als een speciaal geval van criterion-referenced measurement. (Livingston, 1971).
Ebel benadrukt de beperkingen van criterion-referenced measurement. Bovendien wijst hij op de betrekkelijkheid van het begrip "mastery". Ebel stelt voor om de norm-referenced measurement te verbeteren door de inhoudelijke kant beter te ontwikkelen.
"The use of criterion-referenced measurement cannot be expected to improve significantly our evaluations of educational achievement" (Ebel, 1970, 8)
Wijnen tracht aan te tonen dat de principes die ten grondslag liggen aan mastery learning en selektief onderwijs, fundamenteel

vergelijkbaar zijn. Op grond van deze redenering komt hij tot de voorlopige konklusie dat de klassieke testtheorie toepasbaar is voor prestatiemetingen in beide situaties (Wijnen, 1973, 7).
Hoewel de diskussie nog weinig is uitgekristalliseerd en elk standpunt als een zeer voorlopig standpunt moet worden beschouwd, lijkt het zinvol om enkele benaderingen naast elkaar te zetten.
Aan de hand van enkele centrale begrippen uit de psychometrie zal in het volgende de onderscheiding criterion-referenced en norm-referenced measurement verder worden uitgewerkt.
Allereerst wordt hierbij gebruik gemaakt van gegevens uit de literatuur van die auteurs (o.a. Popham en Husek, Warries, van Bockel, Nuy) die op zoek zijn naar een nieuwe psychometrische benadering voor criterion-referenced measurement.
Vervolgens wordt aandacht besteed aan de vraag of het wel noodzakelijk en wenselijk is om alternatieve theorieën, methoden en technieken te ontwikkelen voor prestatiemetingen in het kader van mastery learning.

3.2. Criterion-referenced measurement: psychometrische aspekten.

3.2.1. Variabiliteit.

Voorstanders van een criterion-referenced approach wijzen op het centrale verschil tussen absoluut en relatief meten: de skorevariantie. Omdat de betekenis van een norm-referenced skore is gebaseerd op de relatieve positie van deze skore in vergelijking met andere skores, is het noodzakelijk dat het skorepatroon een grote mate van variantie vertoont: hoe meer variantie hoe beter.
De toets moet sterk differentiëren tussen relatief goede en minder goede leerlingen. Variabiliteit is daarentegen géén noodzakelijke voorwaarde voor een goede criterion-referenced test. Immers de betekenis van de skore is niet afhankelijk van de vergelijking met andere skores. De individuele skore wordt geinterpreteerd in relatie tot het kriterium.
"Het punt waar alles om draait is het feit dat variantie van de testskore bij relatief meten centraal staat en bij absoluut meten niet.
Statistische technieken voor itemselektie en het bepalen van de betrouwbaarheid en validiteit van studietoetsen zijn alle gebaseerd op de variantie van toetsskores.
Voor criterion-referenced test moeten dus andere psychometrische analysetechnieken worden gebruikt.
Alternatieve psychometrische analysetechnieken zijn tot nu echter nauwelijks ontwikkeld". (Nuy, 1973, 660).

3.2.2. Item analyse

Bij itemanalyse kan men een onderscheid maken tussen direkte en indirekte analyse.
Men geeft dit verschil ook wel aan met de begrippen item konstruktie en item analyse. Een direkte analyse vindt plaats tijdens en direkt na de konstruktie van een item, dus voordat de opgaven aan een groep leerlingen zijn voorgelegd. De toetskonstrukteur beoordeelt bij direkte analyse de kwaliteit van de items vooral t.a.v. inhoudsvaliditeit, formuleringseisen en objektiviteit (juistheid)
Een indirekte analyse vindt plaats nadat de opgaven aan

een grote groep leerlingen zijn voorgelegd en men de beschikking heeft over een groot aantal item- en toetsskores. Via statistische bewerkingen krijgt men dan informatie over de psychometrische kwaliteiten van de items en de toets als geheel: p-waarden, a-waarden, D-waarden en indices voor betrouwbaarheid en validiteit. In het algemeen kan worden gezegd dat bij norm-referenced measurement de nadruk is komen te liggen op indirekte analyse terwijl bij criterion-referenced measurement de direkte analyse centraal staat.
Het kenmerkende verschil bij de direkte analyse (itemkonstruktie) van criterion-referenced en norm-referenced tests ligt in de aanpak van de item schrijver (Popham en Husek, 1969, 4).
De item schrijver bij norm referenced measurement zoekt naar items die maximaal differentiëren tussen leerlingen. Hij beoordeelt de items vooral in termen van moeilijkheidsgraad "deze is te gemakkelijk", "die is te moeilijk" etc. Hij probeert een zo groot mogelijke skorevariantie te bereiken. Bij het maken van items voor criterion-referenced tests laat men zich door andere overwegingen leiden. De centrale vraag is hier: is het item een juiste afspiegeling van de expliciet onderwezen stof? M.a.w. het kriterium is bepalend bij de beoordeling van de items, ongeacht p- en D-waarden. Ook bij de indirekte analyse spelen verschillende overwegingen een rol bij norm-referenced en criterion-referenced measurement. Bij het samenstellen van selektieve toetsen moeten volgens van Bockel de volgende kriteria bij de item selektie worden aangelegd: a. goed differentiërende items
b. hoge item-toetskorrelatie.
Voor een "mastery test" noemt van Bockel (in navolging van Warries) 2 voorwaarden voor een goed item:

a. voordat onderricht is gegeven in de leerstof waarop het item betrekking heeft, moet dit item door de meeste leerlingen niet korrekt beantwoord kunnen worden. D.w.z. een lage p-waarde vóór de instruktie.
b. Nadat het betreffende onderricht heeft plaatsgehad moeten bijna alle leerlingen in staat zijn het item goed te beantwoorden. D.w.z. een hoge p-waarde na instruktie.

Bij de selektie van de items kan dan gebruik worden gemaakt van de prétoets-posttoets methode van Cox en Vargas (Nuy, 1973, 667).
Bij de prétoets-posttoets methode vergelijkt men de item skores van geinstrueerde en niet-geinstrueerde groepen. Alleen die items zijn geschikt voor opname in de toets, waarvan het verschil in p-waarde voor en na de instruktie groot genoeg is.
Het percentage leerlingen dat het item op de posttoets goed maakt minus het percentage leerlingen dat het item op de prétoets goed maakt wordt de Dpp-waarde genoemd. (D=diskriminatie-index, pp=prétoets-posttoets).
Enkele problemen die zich voordoen bij het selekteren van items volgens de prétoets-posttoets procedure:

- Behalve bij zeer specifieke training is het meestal moeilijk een niet-geïnstrueerde groep te vinden; met

name in het onderwijs zijn vaak voor de instruktie al delen van de leerstof geleerd.
- Het is bijna altijd onmogelijk om faktoren, die van invloed zouden kunnen zijn op het verschil in p-waarden tussen een niet- en wel-geïnstrueerde groep, zoals leeftijd, intelligentie e.a. onder kontrole houden.
- Bovendien doet zich een ander probleem voor bij de interpretatie: wanneer geen verschil tussen de p-waarden van de beide afnamen wordt gevonden kan dit betekenen dat het item slecht is, maar ook, dat de instruktie wat betreft het in 't item behandelde onderwerp gefaald heeft.
- Tenslotte is het onmogelijk (wanneer dezelfde groep voor en na de instruktie getest wordt) de invloed van de prétest op de prestatie op de posttest te kontroleren (Mastery Learning, 1972).

3.2.3. Validiteit

Procedures voor de validering van norm-referenced toetsen zijn voornamelijk gebaseerd op korrelatie-berekeningen en dus op variabiliteit. Voor absoluut meten zijn de resultaten van deze procedures bruikbaar als ze positief uitvallen.
Omgekeerd kan men niet vaststellen of de toets ongeschikt is wanneer de resultaten negatief zijn.
Criterion-referenced toetsen worden primair gevalideerd in termen van adekwate representatie van het kriterium.
De nadruk valt dus op inhoudsvaliditeit (Popham en Husek 1969, 7)
Nuy omschrijft het verschil tussen norm-referend- en kriterion-referenced measurement m.b.t. de validiteitsvraag ongeveer als volgt: bij relatief meten staat de begripsvaliditeit centraal en gaat het om meer komplexe doelstellingen op langere termijn.
Bij absoluut meten (diagnostische toetsing) staat de inhoudsvaliditeit centraal en gaat het om enkelvoudiger doelstellingen op kortere termijn (Nuy, 1973, 653),(Nuy, 1972, 115).
Het is volgens Nuy erg belangrijk om in teamverband te komen tot een nauwkeurige formulering van doelstellingen in termen van konkrete gedragingen.
Hiervoor kan b.v. gebruik gemaakt worden van Blooms model van leertaakanalyse of van de techniek van regressieve leertaakanalyse van Gagné. (Nuy, 1972, 40 e.v.).
Dit terrein blijkt echter vol te liggen met voetangels en klemmen. Ebel heeft b.v. nogal wat bedenkingen t.a.v. de mogelijkheid om te komen tot kriteriummaten: ze vereisen een zeer gedetailleerde specifikatie van doeleinden of uitkomsten, het is niet realistisch dit te verwachten en het is onpraktisch in gebruik.
De formulering zal meer tijd en moeite kosten dan dit waard is en het gebruik er van zal effektief onderwijs eerder onderdrukken dan stimuleren (Ebel, 1970, 3).
Wijnen vestigt de aandacht op de kloof die er bestaat tussen:
a. taxonomieën en klassifikaties voor onderwijsdoelstellingen en

b. de empirische verifikatie en de praktische bruikbaarheid van deze taxonomieën en klassifikaties in het onderwijs.

Het feit dat een leerkracht een bepaalde klassifikatie kan hanteren bij een leertaakanalyse is nog geen bewijs voor de bruikbaarheid er van.
Bruikbare klassifikaties moeten tenminste vertaalbaar zijn in konkrete aanwijzingen of adviezen voor leerkrachten en leerlingen. Het is nog maar de vraag of informatie in termen van kennis, inzicht, analyse etc. of in termen van signaal-, concept- en principe- leren, kan worden gebruikt voor de verbetering van het onderwijs leerproces. Bovendien dient er nog veel empirisch onderzoek in de onderwijs-situatie te worden gedaan naar de juistheid van de voorgestelde logische klassifikaties (Wijnen, 1973, 4). Hierbij is nog afgezien van andere mogelijke bezwaren t.a.v. het formuleren van specifieke doelen in termen van observerbaar gedrag.

3.2.4. Betrouwbaarheid

De betrouwbaarheidsmaten in de klassieke testtheorie kunnen onder bepaalde voorwaarden wel worden toegepast bij criterion-referenced tests, nl. wanneer de toets wordt gebruikt om één bepaalde dimensie te meten, b.v. een specifiek instruktiedoel.
Popham en Husek wijzen er op dat het beslist niet zo is dat deze indices (gemiddelde inter-item korrelatie en test-hertest betrouwbaarheid) niet gebruikt zouden kunnen worden bij het bepalen van de betrouwbaarheid van de toets. Maar het is bij criterion-referenced toetsen heel goed mogelijk dat de toets zéér consistent is zonder dat dit tot uitdrukking komt in de indices die gebaseerd zijn op variabiliteit. Bij criterion-referenced measurement is het immers niet de bedoeling om te differentiëren binnen de groep maar om na te gaan in hoeverre de in gedragstermen geoperationaliseerde doelstellingen bereikt zijn. Bij toepassing van mastery learning strategieën is het zelfs de bedoeling dat alle leerlingen de maximale skore (het kriterium) bereiken. (Of een bepaald konstant percentage van de maximale skore).
De skorevariantie is dan nul. Dit resulteert dan in een homogeniteitsindex van 0/0. In deze specifieke situatie is de betrouwbaarheid dus niet definieerbaar (Livingston, 1971, 18). Popham en Husek geven enkele zeer voorlopige suggesties voor de richting waarin kan worden gezocht voor het bepalen van alternatieve indices voor de betrouwbaarheid van criterion-referenced tests. In plaats van een interne consistentie-coëfficiënt zou misschien een vergelijkbare maat kunnen worden berekend d.m.v. een methode die grotere temporele eenheden omvat b.v. door pré- en postskores als onderdeel van hetzelfde fenomeen op te vatten. Voorwaarde hiervoor is weer dat de toets een operationalisatie is van slechts een dimensie.
Als in een toets verschillende doelstellingen vertegenwoordigd zijn moeten de groepen items als subtests worden beschouwd. Bij de stabiliteitscoëfficiënt doet zich de-

zelfde moeilijkheid voor bij criterion-referenced measurement bij gebrek aan skorevariantie.
Misschien zou de stabiliteit van de skores kunnen worden aangegeven in termen van een betrouwbaarheidsinterval om de gerealiseerde skore.
Een meer konkreet voorstel voor het bepalen van een betrouwbaarheidscoëfficiënt voor criterion referenced test wordt gedaan door Livingston. Het betreft hier een poging om met gebruikmaking van het klassieke model van testtheorie een criterion-referenced betrouwbaarheidstheorie te ontwikkelen die ook gebruikt kan worden in situaties waarin de skorevariantie gering of zelfs afwezig is.
Livingston zoekt de oplossing door substitutie: in plaats van de mean neemt Livingston het kriterium als referentiepunt bij de berekeningen van de variantie en de betrouwbaarheid.
"Consider the basic difference between norm-referenced measurement. When we use norm-referenced measures, we want to know how far a student's score deviates from the group mean. When we use criterion-referenced measures, we want to know how far his score deviates from a fixed standard, the criterion score. Therefore, each concept based on deviations from the mean score will be a corresponding concept based on deviations from the criterion score. The result of this substitution will be a generalized form of the classical theory of reliability which will include norm-referenced reliability as a special case: the case or which the mean score and the criterion score are equal" (Livingston, 1971, 14).
Het probleem bij deze werkwijze is echter dat de betrouwbaarheidscoëfficiënt van de toets toeneemt naarmate het gekozen kriterium verder af ligt van het gemiddelde terwijl de standaard meetfout gelijk blijft (Harris, 1972, 29). M.a.w. in die gevallen waarin mean en kriterium niet samenvallen levert Livingston's betrouwbaarheidscoëfficiënt een hogere waarde op dan de traditionele betrouwbaarheidscoëfficiënt, maar deze hogere waarde impliceert niet een betere bepaling of een werkelijke skore beneden of boven de kriterium-waarde ligt.
Nog anders gezegd: de hogere Livingston coëfficiënt suggereert ten onrechte een kleiner betrouwbaarheidsinterval. Livingston's coëfficiënt kan dus niet op dezelfde wijze worden gebruikt en geinterpreteerd als de konventionele betrouwbaarheidsmaten.
Livingston claimt dat zijn theorie van betrouwbaarheid toepasbaar is in die situaties waarin geen skore-variantie optreedt d.w.z. wanneer alle leerlingen hetzelfde nivo van beheersing bereiken.
Shavelson, Block en Ravitch stellen hier tegenover dat ook in die gevallen waarbij slechts kleine verschillen tussen de skores optreden m.b.v. de gangbare procedures de betrouwbaarheid en de standaard-meetfout kan worden berekend. En dat op grond van deze informatie een betrouwbaarheidsinterval kan worden gekonstrueerd. (Shavelson e.a. 1972, 135).

4. RELATIEF METEN BIJ MASTERY TESTING

4.1. Afwezigheid van skorevariantie bij mastery testing; een juist uitgangspunt?

Uit het voorgaande is gebleken dat diverse auteurs van mening zijn dat bij het toetsen van prestaties in het kader van mastery learning geen gebruik gemaakt kan worden van de analyse-technieken uit de klassieke testtheorie.
Uitgangspunt en vooronderstelling bij deze redenering was dat bij mastery testing de skore-variantie afwezig (of te gering) is om de gebruikelijke berekeningen, voor het bepalen van de betrouwbaarheid en validiteit etc. te kunnen uitvoeren.
Men redeneert als volgt: als de instruktie geslaagd is (d.w.z. dat alle leerlingen tot beheersing komen) is de skorevariantie nul en is het niet zinvol om de konventionele psychometrische analysetechnieken, die gebaseerd zijn op skorevariantie, toe te passen. De betrouwbaarheid blijft in dit geval ondefinieerbaar.
Nu kan men zich afvragen of de genoemde vooronderstelling wel opgaat. Is het wel juist om te veronderstellen dat bij toepassing van mastery learning strategieën alle leerlingen de maximale skore bereiken. Is het wel efficiënt om het tijdstip van toetsing zó te kiezen dat bijna alle leerlingen tot beheersing zijn gekomen?
Valt er dan eigenlijk nog wel wat te toetsen? Daarmee is tevens de vraag gesteld of het wel noodzakelijk is om een alternatieve meettechnologie te ontwikkelen. Wanneer deze vragen ontkennend moeten worden beantwoord kan het verschil tussen criterion-referenced testing en norm-referenced testing niet langer in termen van aanwezigheid of afwezigheid van skorevariantie worden beschreven. Wanneer men de noodzakelijkheid en/of wenselijkheid van differentiatie in het skorepatroon bij mastery testing zou kunnen aantonen, dan kan men stellen dat gebruik gemaakt kan worden van de begrippen en technieken uit de klassieke testtheorie. (Uiteraard wanneer tevens aan de voorwaarden van het klassieke meetmodel is voldaan). Ebel vindt het een fundamenteel onjuist uitgangspunt om van alle leerlingen een zelfde nivo van beheersing te verlangen.
Elk kriterium voor mastery is noodzakelijkerwijs arbitrair en onvolmaakt. De beperkingen van criterion-referenced measures vat Ebel als volgt samen:

"a. They do not tell us all we need to know about achievement
b. They are difficult to obtain on a sound basis.
c. They are necessary for only a smal fraction of important educational achievements" (Ebel, 1970, 8).

Ook Wijnen is niet overtuigd van de noodzaak en de bruikbaarheid van een alternatieve meettechnologie op dit moment (Wijnen, 1973, 9).
De suggesties van Wijnen komen in de volgende paragrafen aan de orde gevolgd door enkele kritische opmerkingen.

4.2. Fundamentele overeenkomst: mastery learning en selektief onderwijs.

Wijnen stelt in zijn lezing, op het International Symposium on Educational Testing in Den Haag, dat de klassieke testtheorie bruikbaar is bij mastery learning: "If mastery learning, based on formative evaluation should be an alternative to selective

education, which arguments justify in that case the statement, that we need besides classical testtheory, a new theory that fits in the requirement of mastery learning.
My statement is, that classical test theory can serve mastery learning as well as selective education" (Wijnen, 1973, 8).
De argumentatie van Wijnen luidt ongeveer als volgt: er zijn twee kontrasterende benaderingen in de programmering van onderwijs-leerprocessen:

1. programma's waarbij de hoeveelheid beschikbare tijd gelijk is voor alle studenten met als gevolg: individuele verschillen in prestatie (tijd konstant, prestatie variabel)
2. programma's waarbij de vereiste prestatie voor alle studenten gelijk is, met als gevolg: individuele verschillen in tijd. (prestatie konstant, tijd variabel)

De eerste benadering treft men aan in een selektieve onderwijs-leersituatie, terwijl de tweede benadering overeenkomt met mastery learning. De diskussie rondom norm-referenced evaluatie versus criterion-referenced evaluatie is gerelateerd aan deze twee vormen van programmering.
Hoewel men deze beide evaluatievormen in logisch opzicht kan onderscheiden, hoeft dit nog geen pleidooi in te houden voor het ontwikkelen van een nieuwe meettechnologie.
Er zijn namelijk fundamentele overeenkomsten in het denken over beide situaties aan te wijzen.
In een selektieve onderwijssituatie is de vastgestelde hoeveelheid tijd goed gekozen als de gemiddelde prestatie van de groep zo dicht mogelijk ligt bij de doelstellingen van het onderwijs. In dat geval kan men zeggen dat de totale hoeveelheid verspilde tijd zo gering mogelijk is.
De som van wat te veel geleerd is en wat te weinig geleerd is, in relatie tot de doelstellingen, is dan zo klein mogelijk.
In een Mastery Learning situatie kan men de dingen in ongeveer dezelfde termen beschrijven.
Het vastgestelde kriterium is goed gekozen als de totale som van de verschillen tussen studenten m.b.t. de benodigde tijd zo klein mogelijk is. Als er geen evenwicht is tussen studenten die te snel door het programma gaan en studenten die te langzaam studeren, is er iets verkeerd met het vastgestelde kriterium of met de gegeven adviezen.
Als de totale hoeveelheid tijd die de snelle studenten hebben overgehouden aanzienlijk meer is dan de totale hoeveelheid tijd die de trage studenten te kort komen moet het kriterium opnieuw worden geformuleerd. Essentieel in bovenstaande redenering lijkt mij dat, volgens de interpretatie van Wijnen, bij mastery learning het kriterium niet als een absoluut doel moet worden gezien, ongeacht de gemiddelde prestatie van de groep.
Het kriterium moet a.h.w. telkens worden "bijgesteld" afhankelijk van de gemiddelde groepsprestatie.
Tussentijdse relatieve prestatiemetingen kunnen informatie verschaffen omtrent de verschillen in tijd, tussen de studenten onderling, om het gestelde doel te kunnen bereiken, mits men mag uitgaan van een lineaire relatie tussen benodigde tijd en prestatie. Als aan deze vooronderstelling is voldaan, kan men de totale hoeveelheid benodigde tijd om het kriterium te bereiken voorspellen op basis van prestatiemetingen in de begin-fase van het onderwijsleerproces (Block, 1971, 39).

Als b.v. een instruktieprogramma 800 eenheden omvat en een student heeft 10 uur nodig om tot beheersing van de eerste 100 eenheden te komen, dan mag men (gegeven de lineaire relatie) aannemen, dat de totale tijd om het kriterium te bereiken 80 uur zal zijn.
Wijnen stelt dat in een selektieve onderwijsleersituatie deze informatie kan worden gebruikt om de juistheid van de gefixeerde hoeveelheid tijd te bepalen. In een mastery learning situatie kan deze informatie worden gebruikt om de juistheid van het gekozen kriterium te bepalen.
Wanneer men bovenstaande vergelijking aksepteert, inklusief de veronderstelde lineaire relatie tussen prestatie en tijd, moet men toegeven dat relatieve procedures juist bijzonder geschikt zijn bij mastery testing. Dan vervalt de noodzaak om een nieuwe meettheorie te ontwikkelen.
Dit betekent volgens Wijnen echter niet dat elke toets voor elk doel gebruikt kan worden. In het volgende wordt hierop ingegaan aan de hand van enkele modellen van ideaaltypische studietoetsen voor formatieve evaluatie.

4.3. Ideaaltypische studietoetsen.

Men kan in formatieve evaluatie een onderscheiding aanbrengen:

a. formatieve evaluatie t.b.v. het onderwijzen
b. formatieve evaluatie t.b.v. het leren

"Formative evaluation in support of teaching ought to be conceptually distinguished from formative evaluation in support of learning. It hardly can be expected that the same test can serve as a tool to reach two complete different objectives.
I am convinced, that a test which has lead to the improving of an educational programm, needs other properties other than a test which has lead to the improvement of study behavior.
This contrast can be ideally illustrated by the following models". (Wijnen, 1973, 7).

(kolommen zijn vragen, rijen zijn personen)

Model I

I I I 0 0 0	- lage variantie van testskores
I I I 0 0 0	- lage item-totaal korrelaties
I I I 0 0 0	- hoge variabiliteit van p-waarden
I I I 0 0 0	- centraal: item variantie
I I I 0 0 0	- doel: operationalisatie van onderwijs-
I I I 0 0 0	effekten

Model I is bijzonder geschikt voor evaluatie ter verbetering van het onderwijs. De leerkracht kan hiervan veel leren over zijn eigen kapaciteiten (omgekeerde diagnostische evaluatie) M.n. voor het opsporen van punten die nog niet voldoende of niet op de juiste wijze onderwezen zijn. (voortgangstoets) (Wijnen, 1971, 76).

Model II

I I I I I I	- hoge variantie van testskores
I I I I I I	- hoge item-totaal korrelaties
I I I I I I	- lage variabiliteit van p-waarden
0 0 0 0 0 0	- centraal: variantie van personenskores
0 0 0 0 0 0	- doel: operationalisatie van persoons-
0 0 0 0 0 0	- effekten.

Model II is bijzonder geschikt voor verbetering van het leerproces bij mastery learning.
De uitslag van deze toets geeft de student informatie over het resultaat van zijn studiemethoden.
Zeer belangrijk bij beide vormen van formatieve evaluatie is de mogelijkheid van het vertalen van de toetsuitslag in een remediërende onderwijs- en leerstrategie.
Wijnen stelt voor om de toetsen te ontwikkelen in relatie tot eenheden die geisoleerd kunnen worden in het onderwijs programma b.v.: hoofdstukken of topics.
Het is waarschijnlijk zo dat docenten en studenten méér hebben aan informatie in termen van leerstofinhouden (gebieden) dan aan de meer abstrakte klassifikaties zoals die worden gehanteerd door Bloom, Gagné en Mager: b.v. kennis, inzicht, toepassing, signaalleren, principe leren en analyse etc. Deze begrippen laten zich moeilijk vertalen in konkrete leer- en onderwijsstrategieën.
Er bestaat nog steeds een kloof tussen de onderwijskundige praktijk enerzijds en taxonomieën voor onderwijsdoelstellingen en resultaten van analytische methoden anderzijds.
Het volgende model zou misschien o.a. gehanteerd kunnen worden voor researchdoeleinden om de genoemde kloof te overbruggen.

Model III

```
I I I I I I 0 0 0
I I I I I I 0 0 0
I I I I I I 0 0 0
I I I 0 0 0 0 0 0
I I I 0 0 0 0 0 0
I I I 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
```

- ongeveer gelijk aandeel van variantie van vraaggemiddelden en persoonsgemiddelden
- variabiliteit in p-waarden
- kumulatieve homogeniteitsmodel (Guttman-schaal)

Model III kan voor verschillende doeleinden worden gebruikt in het bijzonder voor research.

- empirische verifikatie van voorgestelde logische ordeningen van de leerstof (b.v. Bloom, Gagné)
- verbetering van onderwijzen en leren

Deze toets is gebaseerd op zowel de gegeven instruktie als op de instruktie die nog moet volgen m.a.w. er worden items opgenomen die betrekking hebben op nog te onderwijzen leerstof.

4.4. Kritische opmerkingen.

Hoewel de voorstellen van Wijnen mij wel erg logisch en praktisch voorkomen is het m.i. toch mogelijk om vanuit de theorie achter mastery learning enkele kritische opmerkingen te plaatsen.
Vooral omdat de konceptie van mastery learning als zodanig niet expliciet door Wijnen ter diskussie wordt gesteld.
Dit in tegenstelling tot Ebel die de gedachte van een uniform niveau van mastery voor alle studenten expliciet en principiëel afwijst.
(tenminste voor het merendeel van de gangbare onderwijsdoelstellingen). Onderstaande opmerkingen zijn meer bedoeld als vragen dan als absolute uitspraken.
De kritiek betreft vooral de theoretische vergelijking van

selektief onderwijs en mastery learning. Het impliceert géén ontkenning van de mogelijkheid van praktische toepassing van Wijnen's suggesties.

4.4.1. De argumenten die Wijnen hanteert, om aan te tonen dat de denkwijze bij selektief onderwijs fundamentele overeenkomsten vertoont met de filosofie achter mastery learning zijn gebaseerd op een bepaalde vergelijking tussen beide situaties. Ik heb deze vergelijking voor een groep leerlingen als volgt geschematiseerd:

	tijd	prestatie
selektief onderwijs	konstant	variabel
mastery learning	variabel	konstant

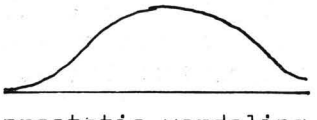
Als men aanneemt dat de leerlingen normaal verdeeld zijn m.b.t. aptitude, kan men stellen dat de prestaties van deze groep onder de kondities van een selektieve onderwijs situatie ook normaal verdeeld zullen zijn (Block, 1971, 6) Deze relatie kan als volgt worden weergegeven:

aptitude

achievement

Wanneer men nu, zoals Wijnen, uitgaat van een lineaire relatie tussen prestatie en tijd per leerling in de groep kan men voor een mastery learning situatie m.b.v. tussentijdse prestatie metingen de verschillen in benodigde tijd voor het gehele onderwijsprogramma bepalen. Wanneer de prestaties normaal verdeeld zijn zal dit ook het geval zijn voor de te verwachten verschillen in benodigde tijd bij een gefixeerd kriterium.

In tekening:



prestatie verdeling voor een groep van leerlingen nadat een deel van het programma is gevolgd

te verwachten verdeling in tijd wanneer men een gefixeerd kriterium stelt voor alle leerlingen van de groep aan het einde van het onderwijsprogramma.

Met bovenstaande voorstelling blijft men m.i. helemaal in de denkwijze van Wijnen. Men kan nu duidelijk 2 vooronderstellingen in deze redenering aangeven.

1. Er is een lineaire relatie tussen prestatie en tijd.
2. Deze relatie geldt voor beide situaties: selektief onderwijs en mastery learning.

De veronderstelde lineaire relatie is op zichzelf al diskutabel. Dit wordt door Wijnen ook opgemerkt.
Maar vooral binnen de kontekst van mastery learning lijkt het mij niet geoorloofd om van een lineaire relatie uit te gaan. De vergelijking (zie schema) tussen beide onderwijssituaties gaat niet op omdat slechts twee relevante variabelen in de vergelijking zijn opgenomen, nl. prestatie en tijd.
Andere zeer belangrijke variabelen die buiten de vergelijking zijn gebleven zijn: aptitude, quality of instruction, ability to understand instruction en de interaktie tussen deze variabelen. De vergelijking en de lineaire relatie geldt bij mastery learning alleen aan het begin van het programma b.v. na een korte periode van uniforme instruktie M.a.w. wanneer de leerlingen nog sterk verschillen in beginsituatie. Het resultaat dat met mastery learning wordt beoogd is niet alleen dat (bijna) alle leerlingen criterion-performance bereiken, maar ook dat de individuele verschillen in tijd steeds meer zullen verdwijnen (Block, 1971, 55)
Hierdoor wordt het m.i. onmogelijk om d.m.v. tussentijdse relatieve metingen de te verwachten spreiding in de tijd te voorspellen. Het is dan niet zinvol om op grond van deze informatie het kriterium te veranderen om tot een meer optimale planning te komen zoals Wijnen voorstelt (Wijnen, 1973, 10).
De verschillen in leertijd zijn vooral bij mastery learning niet stabiel voor een langere instruktieperiode en voor elke leertaak. Er kunnen belangrijke fluktuaties optreden in de leertijd van een bepaalde leerling, juist als gevolg van manipulatie van relevante variabelen zoals kwaliteit van de instruktie etc.

4.4.2. Opvallend is dat Wijnen één ideaaltypische studietoets (model II) geschikt acht voor zowel individuele diagnose (Wijnen, 1973, 7) als voor selektie (Wijnen, 1971, 72, e.v.).
Dit is niet in overeenstemming met zijn opmerking: "It hardly can be expected that the same test can serve as a tool to reach two complete different objectives" (Wijnen, 1973, 7).

4.4.3. Volgens Wijnen dient het kriterium bij mastery learning bepaald te worden in relatie tot de gemiddelde groepsprestatie. Men kan zich afvragen of dit wel in overeenstemming is met de idee van mastery learning.

4.4.4. Misschien zou men kunnen stellen dat Wijnen:

a. de filosofie achter mastery learning enigszins aanpast aan het denken in een selektieve onderwijs situatie.
b. vervolgens konkludeert dat er fundamentele overeenkomsten zijn in beide denkwijzen.
c. op grond van a en b tracht aan te tonen dat de klassieke testtheorie juist heel goed bruikbaar is bij mastery testing.

5. SAMENVATTING EN KONKLUSIES

5.1. Absoluut en relatief meten.

Het onderscheid tussen absoluut en relatief meten kan (in theorie) m.b.t. de skoringsprocedure eenduidig worden beschreven. Formeel bestaat er overeenkomst tussen beide meetmethoden: zowel bij absoluut als bij relatief meten moet een 3-tal beslissingen worden genomen:

a. het bepalen van het referentie punt.
b. het bepalen van de meeteenheid.
c. het kwantificeren van de toelaatbare afstand, uitgedrukt in de gekozen meeteenheid.

Deze 3-fasen kunnen in principe onafhankelijk van elkaar verlopen.
Inhoudelijk is er een duidelijk verschil. Het kenmerkende verschil ligt hierin dat bij relatief meten de skoringsprocedure wordt gekoppeld aan de groepsprestatie, terwijl bij absoluut meten de aktuele prestatie van de groep géén rol speelt.
Bij absoluut meten wordt het kriterium inhoudelijk bepaald in relatie tot de expliciet onderwezen leerstof.
Zowel bij absoluut als bij relatief meten zijn vele konkrete oplossingen denkbaar, afhankelijk van de inhoud die men geeft aan elk van de drie genoemde beslissingen.
Theoretisch (ideaaltypisch) kan men een onderscheid maken tussen absoluut en relatief meten.
Praktisch blijkt een strikt absoluut oordeel niet mogelijk te zijn. "Absolute" prestatienormen zijn in feite generalisaties op basis van voorgaande ervaringen.
M.a.w. in praktische onderwijsleersituaties zijn "absolute" oordelen steeds min of meer relatief.

5.2. Criterion-referenced measurement versus norm-referenced measurement.

Wanneer men de onderscheiding absoluut versus relatief meten plaatst binnen het kader van de diskussie rondom mastery learning en mastery testing, wordt meestal gebruik gemaakt van de begrippen criterion-referenced measurement en norm-referenced measurement. Verschillende auteurs zijn van mening dat de genoemde onderscheiding niet alleen een kwestie van skoringsprocedure is.
Men wijst er op dat deze tweedeling gerelateerd is (of: moet zijn) aan een tweedeling in toetsdoelen, onderwijsleersituaties, onderwijsfilosofieën, meetinstrumenten en meettheorieën.
Enigszins schematiserend kan men dit als volgt samenvatten:

	norm-referenced meas.	criterion-ref. meas.
skoringsprocedure	relatief meten	absoluut meten
toetsdoel	selektie	individuele diagnose
onderwijsleersituatie	selektieve o.l.situatie	mastery learning
onderwijsfilosofie	schaarste filosofie, wedijver, prestatiedrang	mastery learning filosofie
meetinstrument	norm-referenced test o.b.v. skorevariantie	criterion-referenced test
meettheorie	klassieke testtheorie (aptitude measurement)	alternatieve meet-technologie

Bovenstaande schema komt overeen met de suggesties van Nuy, Warries, Popham en Husek e.a.

Als samenvatting van dit literatuur-onderzoek geeft dit schema echter een eenzijdig beeld van de diskussie.
Andere deskundigen (o.a. Ebel, Shavelson e.a., Wijnen) geven de voorkeur aan toepassing van de klassieke testtheorie voor formatieve evaluatie.

- Met betrekking tot het aspekt, toetsdoel, blijkt uit de jarenlange diskussie dat voor selektie zowel van relatief als van absoluut meten gebruik gemaakt kan worden.
  Men kan b.v. zéér streng selekteren m.b.t. absolute meetmethoden.
- Dit betekent tegelijkertijd t.a.v. de aspekten onderwijsleersituatie en onderwijsfilosofie, dat criterion-referenced measurement ook in een selektief onderwijsklimaat kan worden gebruikt. Dit is namelijk geheel afhankelijk van de konkretisering van de skoringsprocedure m.a.w. van de bepaling van het kriterium.
  Omgekeerd wordt door Wijnen gesteld dat relatief meten allerminst in verband hoeft te worden gebracht met een selektieve onderwijssituatie en een filosofie o.b.v. schaarste. (Wijnen, 1971, Ned. Tijdschr. v. Psych.).
- Het ideaaltype studietoets dat bij uitstek geschikt is voor selektiedoeleinden is volgens Wijnen ook bruikbaar als ideaaltype bij de konstruktie van studietoetsen voor mastery testing. (Zie ook model II).
- Voor wat betreft het aspekt meettheorie impliceert de suggestie van Wijnen de toepasbaarheid van de klassieke testtheorie in mastery learning situaties.
- Om de tweedeling, zoals voorgesteld in het schema, nog verder te relativeren kan worden opgemerkt dat het in een bepaalde konkrete situatie mogelijk is dat relatief meten tot dezelfde uitslag leidt als absoluut meten. In dat geval kan men absoluut meten opvatten als een specifiek geval van norm-referenced measurement.

- Vervolgens kan men stellen dat een kriterion-referenced test ook gebruikt kan worden als een norm-referenced test.
- Tenslotte blijkt dat absoluut meten in strikte zin een illusie is en dat dit in de praktijk neerkomt op het willekeurig buiten beschouwing laten van informatie die relevant is voor de bepaling van de norm. (Hofstee, 1973).

5.3. Konklusies

De stand van de diskussie geeft aanleiding tot enkele voorlopige konklusies:

1. Het zoeken naar alternatieve begrippen en methoden voor criterion-referenced measurement blijkt op grote moeilijkheden te stuiten.
   Bovendien wordt de noodzaak hiervan door sommigen sterk in twijfel getrokken.
2. De uitspraken van de diverse auteurs zijn niet altijd onderling vergelijkbaar omdat ze impliciet of expliciet uitgaan van verschillende konkrete situaties en definities.
   Dit lijkt m.n. in de diskussie tussen Warries en Wijnen het geval te zijn.
   Warries heeft voorkeur voor absoluut meten o.b.v. een alternatieve meettechnologie, bij teacher made toetsen voor diagnostisch en evaluerend gebruik in het voortgezet onderwijs.
   Wijnen heeft een zekere voorkeur voor een relatieve meetmethode, gebaseerd op de klassieke testtheorie, zowel voor selektie als voor formatieve evaluatie.
   Hij schijnt daarbij vooral te denken aan professionele toetsen, te gebruiken bij het universitair onderwijs en bij research. (empirische verifikatie van klassifikatie systemen voor onderwijsdoelen).
   Het begrip absoluut meten wordt in verschillende betekenissen gebruikt: pragmatisch, in termen van leerstof (Warries) of in meer strikte zin (Hofstee).
3. Het lijkt vooralsnog niet zinvol om een definitieve keuze te bepalen tussen absoluut en relatief meten. Het betreft hier twee meetprocedures die afhankelijk van de situatie kunnen worden toegepast. Op dit moment is echter niet precies aan te geven voor welke situaties de afzonderlijk procedures het meest geschikt zijn. Wel dient men zich bij absoluut meten bewust te zijn van de betrekkelijkheid van het begrip "absoluut" in dit verband. Voor o.a. selektiedoeleinden zijn grote bezwaren aan absoluut meten verbonden.
4. De meest praktische oplossing voor dit moment lijkt mij de verdere ontwikkeling van de klassieke testtheorie, ook voor formatieve evaluatie.
   In het bijzonder bij de ontwikkeling van professionele toetsen. Men mag waarschijnlijk aannemen dat ondanks toepassing van mastery learning strategieën, bij formatieve evaluatie voldoende skorevariantie overblijft om de gebruikelijke analysetechnieken te kunnen hanteren.
   Dit zal met name het geval zijn als aan de volgende voorwaarden wordt voldaan:

   a. betrekkelijk heterogene groepen

b. ook items worden opgenomen die behoren tot nog te onderwijzen leerstof
c. het tijdstip van toetsing niet zolang wordt uitgesteld totdat (bijna) alle leerlingen het kriterium hebben bereikt.

6. LITERATUUR

Bakker, R. — Lot en daad, geluk en rede in het Griekse denken. Utrecht, Bijleveld, 1969.

Block, J.H. (ed.) — Mastery Learning. New York, Holt, Rinehart and Winston, 1971.

Bockel, C.A. van — Itemselektie bij Mastery Testing. Amsterdam, R.I.T.P., mei, 1971.

Carroll, J.B.A. — A model of school learning. Teachers College Record, 64 (1963) no.8, 723-734.

Ebel, R.L. — Measuring Educational Achievement. Englewood Cliffs, New Yersey, Prentice Hall, 1965.

Ebel, R.L. — Some Limitations of Criterion-Referenced Measurement. Amsterdam, 1970. Lezing R.I.T.P., augustus.

Emrick, J.A. — An Evaluation Model for Mastery Testing. Journal of Educational Measurement, 8(1971) no.4, 320-326.

Glaser, R. — Instructional Technology and the Measurement of Learning Outcomes; some Questions. American Psychologist, 18(1963) no.8, 519-521.

Groot, A.D. de en R.F. van Naerssen — Studietoetsen, construeren, afnemen, analyseren. Den Haag, Mouton, 1969.

Harris, C.W. — An Interpretation of Livingston's Reliability Coëfficiënt for Criterion-Referenced Tests. Journal of Educational Measurement, 9(1972) no.1, 27-30.

Hofstee, W.K.B. — Een alternatief voor normhandhaving bij toetsen. Nederlands Tijdschrift voor de Psychologie, 28(1973) no.8/9, 215-227.

Lagerweij, N.A.J., J.J. Peters e.a. — Differentiatie in het onderwijs: enige aspekten nader belicht. Info, 4(1973) no.6, 221-264.

Livingston, S.A. — Criterion-Referenced Applications of Classical Test Theory. Journal of Educational Measurement, 9(1971) no. 1, 13-26.

Livingston, S.A. — Reply to Harris. An Interpretation of Livingston's Reliability Coëfficiënt for Criterion-Referenced Tests. Journal of Educational Measurement, 9(1972) no.1, 31-32.

Mastery Learning — Amsterdam, 1972. Seminariumverslag, R.I.T.P., augustus.

Meuwese, W — Onderwijsresearch. Utrecht enz., Spectrum, 1970. Aula-boeken, 437.

Nuy, M.J.G. — Diagnostische Toetsen. 's Hertogenbosch, Malmberg, 1972. Bouwstenen voor experimenten, Reeks van het Kath. Ped. Centrum.

Nuy, M.J.G. — A general theoretical framework for individualized instruction. Pedagogische Studiën, 49(1972) no.4, 167-180.

Nuy, M.J.G. — Psychometrische aspecten van criterion-referenced tests. Nederlands Tijdschrift voor de Psychologie, 27(1972) no.10, 648-678.

Popham, W.J. and T.R.Husek — Implications of Criterion-Referenced Measurement. Journal of Educational Measurement, 6(1969) no.1, 1-9.

Shavelson, R.J., J.H. Block and M.M.Ravitch — Criterion-Referenced Testing: Comments on Reliability. Journal of Educational Measurement, 9(1972) no.2, 133-137.

Störig, H.J. — Geschiedenis van de filosofie I. Utrecht enz., Het Spectrum, 1966.Prisma-boeken, 410.

Thelen, H.A. — Dynamics of Groups at Work. Chicago enz. The University of Chicago Press, 1970.

Warries, E. — Het relatief meten van leerprestaties in het onderwijs. Nederlands Tijdschrift voor de Psychologie, 25(1970) 429-439.

Warries, E. — Drie redenen om te toetsen in het onderwijs. Pedagogische Studiën, 48(1971) no.4, 152-161.

Warries, E. — Het relatief meten van leerprestaties in het onderwijs: dupliek. Nederlands Tijdschrift voor de Psychologie, 26(1971) no.9/10, 596-599.

Wijnen, W.H.F.W. — Betrekkelijkheid van de bezwaren tegen relatief meten. Naar aanleiding van Warries: Het relatief meten van leerprestaties in het onderwijs. Ned.T.Ps. 25(1970) 429-439. Nederlands Tijdschrift voor de Psychologie, 26(1971) no.2, 135-140.

Wijnen, W.H.F.W. — Onder of boven de maat.
Amsterdam, Swets & Zeitlinger, 1971.
Een methode voor het bepalen van de grens voldoende/onvoldoende bij studietoetsen.

Wijnen, W.H.F.W. — Formative Evaluation and Educational Testing.
Groningen, Centrum Onderzoek Wetenschappelijk Onderwijs, 1973.
Paper, aangeboden op de International Conference on Educational Testing, juli.

Groningen, september 1973.

Jan Terwel.